
بسم الله الرحمن الرحيم

پردیس علوم
دانشکده ریاضی، آمار، علوم کامپیوتر

طراحی یک دادگان دست‌نوشته برون‌خط پیوسته فارسی

نگارنده:

نادیا قبادی پاشا

استاد راهنما:

دکتر باقر باباعلی

پروژه برای دریافت درجه کارشناسی
در رشته علوم کامپیوتر
تیر ۹۵

سپاس‌گزاری

سپاس فراوان از زحمات جناب آقای دکتر باباعلی

فهرست مطالب

چکیده

در حوزه پردازش متن، یکی از مباحث اساسی شناسایی متون است که یکی از زمینه فعال آن شناسایی دست‌نوشته برای زبان‌های مختلف می‌باشد. در زمینه شناسایی دست‌نوشته در حوزه‌های زبان‌هایی که از نوشتار لاتین استفاده می‌شود، تحقیقات زیادی صورت گرفته که نتایج آن در قالب محصولات مختلفی در بازار عرضه شده است. یکی از ملزومات اساسی تحقیقات در این زمینه، وجود دادگان‌های دست‌نوشته است که برای زبان‌های لاتین به تنوع موجود است. ولی برای حوزه زبان فارسی متأسفانه همچنین دادگانی با کاربرد تشخیص دست‌نوشته پیوسته موجود نیست و یا قابل دسترسی برای محققین این حوزه نیست. دادگان‌های دست‌نوشته موجود، شامل اعداد و یا حروف گسسته و یا در بهترین حالت کلمات گسسته دست‌نویس می‌باشد. در این پروژه هدف، طراحی یک دادگان دست‌نوشته برون‌خط پیوسته فارسی می‌باشد که از لحاظ متن دست‌نوشته، نوع قلم، تعداد نویسنده، جنسیت و میزان تحصیلاتشان، چپ یا راست دست بودن و سبک نوشتاری از جامعیت کافی برخوردار باشد.

فصل ۱

مقدمه

زبان فارسی میراث مشترک همه اقوام ایرانی است که در فلات پهناور ایران زمین سالیان سال زیسته‌اند. این زبان فاخر سال‌های سال زبان علم و زبان ادبیات و زبان تاریخ و زبان دین و فلسفه بوده است وبخشی از تمدن بشری به این زبان خلق شده و محفوظ است بنابراین حفظ و تواناسازی زبان فارسی نه تنها وظیفه ملی بلکه وظیفه ای جهانی و بشری است. که به حق در سند چشم انداز و نقشه جامع علمی کشور به آن توجه شده و به عنوان زبان علم مورد تاکید قرار گرفته‌است. در روزگار ما که دنیای ارتباطات محسوب می‌شود فضاهای مجازی از گسترده ترین و قابل دسترس‌ترین فضاهای ارتباطی است و طبعا ارتقا جایگاه زبان فارسی در این فضای ارتباطی به حفظ و توسعه آن کمک بزرگی خواهد رساند. بر این اساس پردازش خط و زبان برای زبان های مختلف کشورهای دنیا جزو تحقیقات ضروری حوزه فناوری و اطلاعات در آمده است. از آنجا که اجزا اصلی زبان فارسی با زبان های غربی به طور کلی متفاوت می باشد، روش ها و الگوریتم‌های طراحی شده برای پردازش در این زبان‌ها حداقل

به طور مستقیم برای زبان فارسی قابل استفاده نمی‌باشد. لذا تمرکز تحقیقات بر روی پردازش خط و زبان فارسی در ابعاد مختلف آن از جمله پردازش دیجیتالی متون فارسی اعم از چاپی و دست‌نویس، می‌باشد.

۱.۱ شناسایی الگو

شناسایی الگو، شاخه ای از هوش مصنوعی[۱] است که با طبقه‌بندی[۲] و توصیف مشاهدات سروکار دارد.[۳] شناسایی الگو به ما کمک می‌کند داده‌ها (الگوها) را با تکیه بر دانش قبلی یا اطلاعات آماری استخراج شده از الگوها، طبقه‌بندی نماییم. الگوهایی که می بایست طبقه‌بندی شوند، معمولا گروهی از سنجش‌ها یا مشاهدات هستند که مجموعه نقاطی را در یک فضای چند بعدی مناسب تعریف می‌نمایند. یک سیستم شناسایی الگوی کامل متشکل است از یک حسگر که مشاهداتی را که می‌بایست توصیف یا طبقه‌بندی شوند جمع‌آوری می‌نماید، یک سازوکار برای استخراج ویژگی‌ها که اطلاعات عددی یا نمادین را از مشاهدات، محاسبه می‌کند (این اطلاعات عددی را با یک بردار بنام بردار ویژگی‌ها نمایش می‌دهند) و یک نظام طبقه‌بندی یا توصیف که وظیفه اصلی طبقه‌بندی یا توصیف الگوها را با تکیه بر ویژگی‌های استخراج شده عهده‌دار است.

[۱] Artificial Intelligence

[۲] Classification

۲.۱ آنالیز تصویر اسناد

مبحث «آنالیز تصویر اسناد» (دی‌آی‌ای)[۳] از جمله شاخه‌های فعال در شناسایی الگو و پردازش تصاویر می‌باشد و مشتمل بر کلیه مراحل پردازشی است که محتویات یک سند اسکن یا فکس‌شده را به یک فرم الکترونیکی مناسب، تبدیل می‌نماید. تکنیک‌های «دی‌آی‌ای» اجزای مختلف ساختاری سند، یعنی قسمت‌های متنی (پاراگراف‌ها، کلمات، حروف و ...)، قسمت‌های گرافیکی (خطوط، نمادها، نمودارها و ...) و قسمت‌های تصویری (تصاویر موجود در متن) را از یکدیگر تفکیک می‌کنند و پردازش مناسب را بر روی هر دسته از اجزا، اعمال می‌نمایند و نیز با توجه به ارتباط «منطقی» بین اجزای مختلف، نقش هر یک از این اجزا را در سند مشخص می‌سازند. «دی‌آی‌ای» شامل تکنیک‌های «بازشناسی حروف»[۴] می‌باشد. این تکنیک‌ها درمورد اجزایی از تصویر سند که توسط تکنیک‌های تحلیل در آنالیز تصویر اسناد به عنوان متن تشخیص داده می‌شوند، اعمال می‌گردند و تصویر سند را به یک متن قابل ویرایش توسط رایانه تبدیل می‌نمایند.

سیستم های بازشناسی نوری حروف با حذف نقش تایپیست‌ها در فرآیند تبدیل اسناد کاغذی به قالب الکترونیکی، سرعت ورود اطلاعات به رایانه را ده‌ها برابر افزایش می‌دهند و روند انجام این فرایند را به میزان قابل توجهی تسهیل می‌کنند. امروزه بازار مصرف این سیستم‌ها طیف بسیار وسیعی از مؤسسات (شامل مراکز نشر، دانشگاه‌ها، کتابخانه‌ها، بانک‌ها، ادارات پستی، شرکت‌های بیمه، و ...) را دربرمی‌گیرد. و یکی از دشوارترین زمینه‌های شناسایی الگو می‌باشد. اکثر کارهای انجام شده در این زمینه دررابطه با متون لاتین، چینی و ژاپنی بوده است. سیستم نویسه‌خوان نوری فارسی با وجود

[۳] Document Image Analysis (DIA)

[۴] Optical Character Recognition (OCR)

حجم نسبتا وسیع تحقیقات دانشگاهی و نیاز شدید بازار تجاری به آن، هنوز هم از جایگاه مورد نظر فاصلهٔ بسیاری دارد و تاکنون هیچ سیستم نویسه‌خوان نوری کارآمدی که از نظر دقت و کیفیت محیط نرم‌افزاری، قابل مقایسه با سیستم‌های نویسه‌خوان نوری لاتین باشد، عرضه نگردیده‌است. در نتیجه ضرورت انجام تحقیقات بیشتر در زمینهٔ متون فارسی و عربی کاملا احساس می‌شود. به واسطه‌ی وجود تفاوت‌های اساسی بین نحوه‌ی نگارش فارسی و لاتین (نظیر چسبیده بودن حروف کلمه به یکدیگر، تغییر شکل حروف براساس موقعیت نسبی آن در کلمهٔ فارسی، و ...)، امکان اعمال مستقیم رو ش‌های بازشناسی متون لاتین به منظور شناسایی متون فارسی وجود ندارد.

۳.۱ نقش دادگان در سیستم‌های نویسه‌خوان نوری

مهم‌ترین عامل در تشخیص دست‌نوشته وجود یک دادگان استاندارد می‌باشد. الگوریتم‌های متنوعی برای تشخیص دست‌نوشته موجود است. برای مقایسه‌ی عملکرد الگوریتم‌های متفاوت تشخیص دست‌نوشته، این الگوریتم ها باید روی یک دادگان یکسان و استاندارد آزمایش شوند. در زبان‌های انگلیسی و عربی دادگان استاندار و کافی موجود می‌باشد. اما در زبان فارسی یک دادگان دست‌نوشته استاندارد برای ارزشیابی الگوریتم‌های متفاوت موجود نمی‌باشد. در این پروژه، هدف طراحی یک دادگان دست‌نوشته برون‌خط پیوسته فارسی می‌باشد.

فصل ۲

سیستم‌های بازشناسی دست‌نوشته

شناسایی متون چاپی و دست‌نویس، سالیان طولانی موضوع تحقیق و پژوهش گسترده بوده، که در این میان زبان انگلیسی بیشترین سهم از تحقیقات را به خود اختصاص داده‌است. شناسایی متون چاپی انگلیسی، به علت جدا بودن حروف، از مسائل ساده‌تر این حوزه محسوب می‌شود. اما در مورد زبان فارسی، به علت پیوستگی حروف (حتی در متون چاپی)، شناسایی متن هنوز به عنوان یک موضوع نسبتاً دشوار مطرح است. عمده تحقیقات مربوط به زبان فارسی را پژوهش‌های داخلی تشکیل می‌دهند، هرچند پژوهش‌های زبان عربی نیز به اندازه کافی به این موضوع شباهت دارد. دستخط‌های فارسی و عربی دارای اختلافات جزئی هستند. از نقطه نظر بازشناسی، دست‌خط عربی کمی بیشتر پیچیده است، [۶] و [۷] اما شباهت‌های مابین این دو زبان بر تفاوت‌هایشان ارجح بوده و از اهمیت بیشتری برخوردار است.

۱.۲ معرفی سیستم بازشناسی نوری حروف

در چند دهه‌ی گذشته بازشناسی الگوهای نوشتاری شامل حروف، ارقام و دیگر نمادهای متداول در اسناد نوشته‌شده به زبان‌های مختلف، توسط گروه‌های مختلفی از محققین مورد مطالعه و بررسی قرار گرفته‌است. نتیجهٔ این تحقیقات منجر به پیدایش مجموعه‌ای از روش‌های سریع و تا حد زیادی مطمئن موسوم به نویسه‌خوان نوری یا «بازشناسی نوری حروف» به منظور وارد نمودن اطلاعات موجود در اسناد، مدارک، کتاب‌ها و سایر مکتوبات تایپی و حتی دست‌نوشت به داخل رایانه شده‌است. اصطلاح نویسه‌خوان نوری به تکنیک‌هایی اطلاق می‌شود که در تصاویر اسکن یا فکس‌شده، نواحی متنی را تشخیص می‌دهند و سپس این نواحی (تصویری) را به متن قابل ویرایش تبدیل می‌نمایند.

با دستگاهی به نام اسکنر [۱] می‌توان تصویر یک صفحه کاغذ را به صورت یک فایل گرافیکی (تصویری)، به رایانه ارسال و در آن ذخیره نمود. بدین ترتیب کاربر می‌تواند با یک نرم‌افزار مناسب نمایش‌دهنده‌ی تصاویر، تصویر صفحه‌ی‌اسکن‌شده را بر روی نمایشگر رایانه‌ی خود ملاحظه نماید یا آن را چاپ کند؛ اما قادر نخواهد بود که متن موجود در تصویر سند را ویرایش کند یا آن را مورد جستجو قرار دهد. یک نرم‌افزار نویسه‌خوان نوری، تصویر اسکن‌شده را می‌خواند، محتویات آن (شامل متن، خطوط، تصاویر، جداول و ...) را شناسایی می‌نماید، و سپس آن را به یک قالب قابل ویرایش (در واژه‌پردازها) تبدیل می‌کند. امروزه بیشتر دستگاه‌های اسکنر به نرم‌افزارهای نویسه‌خوان نوری مجهز گردیده‌اند و قادرند متن موجود در یک سند اسکن‌شده را تشخیص دهند و آن را با همان نحوه‌ی قالب‌بندی، ستون‌بندی، جدول‌بندی و نوع فونت مطابق با سند کاغذی اصلی، در قالب یک فایل متنی با قالب‌بندی مناسب ذخیره نمایند.

[۱] Scanner

۲.۲ تاریخچه سیستم‌های نویسه‌خوان نوری

از جنبه‌ی تاریخی، سیستم‌های نویسه‌خوان نوری تا کنون سه مرحله‌ی تکاملی را پشت سر گذاشته‌اند.

۱.۲.۲ مرحله‌ی تکوین (از ۱۹۰۰ تا ۱۹۸۰)

ردپای اولیه‌ی اقدامات صورت‌گرفته در زمینه‌ی بازشناسی حروف را در سال‌های اول دهه‌ی ۱۹۰۰ می‌توان یافت و آن زمانی است که «تیورینگ»[۲] دانشمند روسی بر آن بود که به افراد مبتلا به نارسایی‌های بینایی کمک نماید. اولین اختراع‌های ثبت‌شده در این زمینه مربوط به سال‌های ۱۹۲۹ و ۱۹۳۳ میلادی هستند. این سیستم‌ها حروف چاپی را با روش تطابق قالبی[۳] شناسایی می‌کردند؛ به این صورت که ماسک‌های مکانیکی مختلفی از مقابل تصویر حرف عبور می‌کردند (مکانیکی) و نور از یک سو به آن تابانده می‌شد و از سوی دیگر توسط یک آشکارساز نوری دریافت می‌گردید (اپتیکی). وقتی یک انطباق کامل صورت می‌گرفت، نور به آشکارساز می‌رسید و حرف ورودی بازشناسی می‌شد. این اختراع به دلیل فناوری اپتومکانیکی مورد استفاده در آن، کاربردی نبود. تصور دسترسی به دستگاهی برای بازشناسی حروف تا دهه‌ی ۱۹۴۰ میلادی و ظهور رایانه‌های دیجیتال، به صورت یک رویا باقی ماند.

اقدامات اولیه در زمینه‌ی بازشناسی حروف، بر متون چاپی یا مجموعه‌ی کوچکی از حروف و نمادهای دست‌نوشت که به راحتی قابل تشخیص بودند، متمرکز گردیده بود. سیستم‌های بازشناسی

[۲] Tyuring

[۳] Template matching

حروف چاپی که در این مقطع زمانی عرضه شدند، عمدتا از روش تطابق قالبی استفاده می‌نمودند که در آن، تصویر ورودی با مجموعه بزرگی از تصاویر حروف، مورد مقایسه قرار می‌گرفت. در مورد متون دست‌نوشت نیز الگوریتم‌های پردازش تصویر که ویژگی‌های سطح پایین[۴] (ویژگی‌هایی که مستقیما و بدون اعمال هیچ تبدیلی، از تصاویر استخراج می‌شوند) را از تصاویر استخراج می‌کنند، در مورد تصاویر دوسطحی[۵] اعمال می‌شدند تا بردارهای ویژگی استخراج گردند. سپس این بردارهای ویژگی به طبقه‌بندی‌کننده‌های آماری سپرده می‌شدند. در این دوره، تحقیقات موفق اما مقید [۶] (منظور از مقید، مفروض دانستن شرایط و پیش‌فرض‌های خاص برای کاراکترهای ورودی است)، بیشتر بر روی حروف و اعداد لاتین انجام گرفت. با این حال مطالعات چندی نیز بر روی حروف ژاپنی، چینی، عبری، هندی، سیریلیکی، یونانی و عربی در هر دو زمینه‌ی حروف چاپی و دست‌نوشت آغاز گردید. با ظهور صفحات رقومی‌کننده[۷] در دهه‌ی ۱۹۵۰ که قادر به تشخیص مختصات حرکتی نوک یک قلم مخصوص بودند، سیستم‌های نویسه‌خوان نوری تجاری نیز امکان عرضه یافتند. این نوآوری سبب شد که محققان بتوانند در زمینهٔ بازشناسی برخط[۸] حروف دست‌نوشت، فعالیت خود را آغاز نمایند.

۲.۲.۲ مرحله‌ی توسعه (از ۱۹۸۰ تا ۱۹۹۰)

مطالعات صورت گرفته تا قبل از سال ۱۹۸۰ به دلیل فقدان سخت‌افزارهای رایانه‌ای قدرتمند و دستگاه‌های اخذ داده‌ها با مشکل همراه بودند. در این دهه به واسطه‌ی رشد انفجارگونه‌ی فناوری اطلاعات، وضعیت بسیار مناسبی برای تحقیقات مختلف از جمله بازشناسی حروف فراهم گردید.

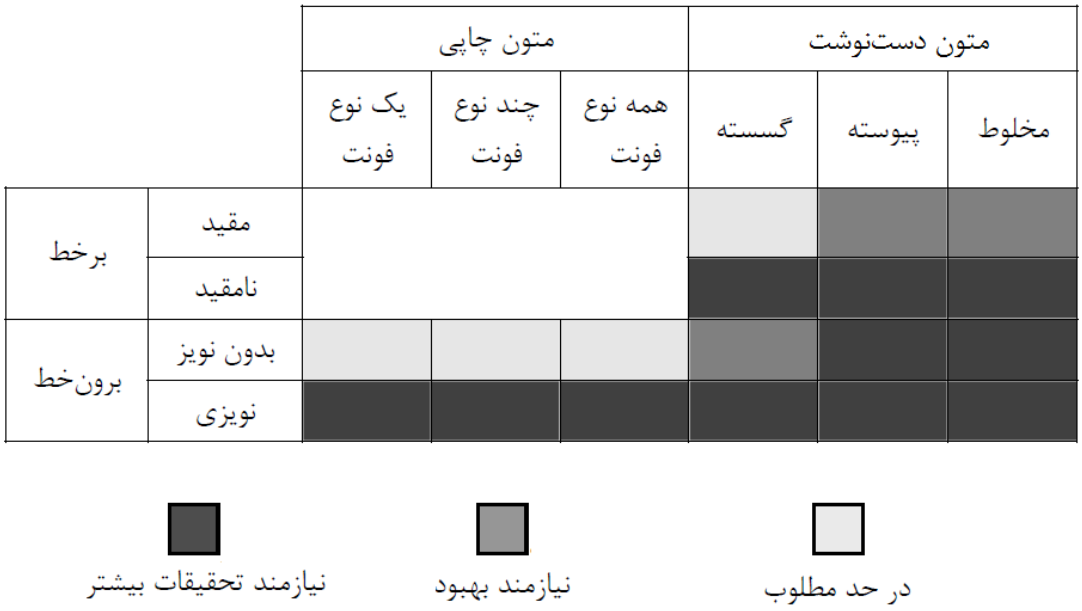
[۴] Low level features
[۵] Binary
[۶] Constrained
[۷] Digitizers
[۸] Online

روش‌های ساختاری به همراه روش‌های آماری در بسیاری از سیستم‌ها استفاده شدند. تحقیقات در زمینه‌ی نویسه‌خوان نوری اساسا به توسعه روش‌های بازشناسی معطوف گردید، بی‌آنکه مسئله استفاده از اطلاعات معناشناختی[۹] به منظور افزایش دقت بازشناسی مورد توجه قرار گیرد. این امر سبب گردید که دقت بازشناسی (نرخ بازشناسی) از یک حد خاص فراتر نرود، که در بسیاری از کاربردهای نویسه‌خوان نوری، قابل قبول نبود.

۳.۲.۲ مرحله‌ی بهبود (از ۱۹۹۰ به بعد)

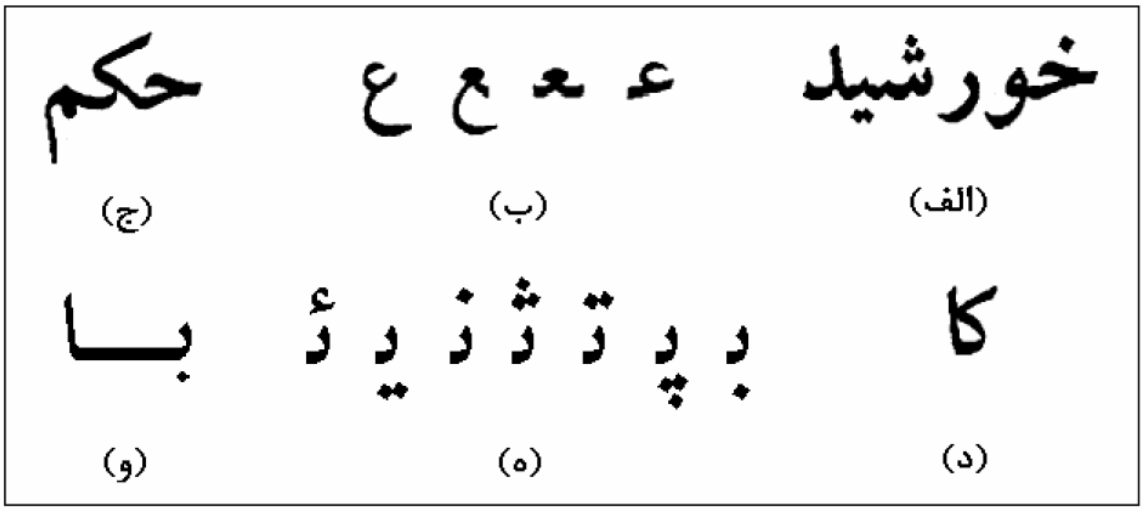
در این مقطع زمانی بود که با تکوین ابزارها و تکنیک‌های پردازشی جدید، پیشرفت واقعی در سیستم‌های نویسه‌خوان نوری محقق گردید. در اوایل دهه‌ی ۹۰، روش‌های پردازش تصویر و بازشناسی الگو با تکنیک‌های کارآمد هوش مصنوعی ادغام گشتند. محققان، الگوریتم‌های پیچیده‌ای را در بازشناسی حروف ابداع نمودند که قادر بودند داده‌های ورودی با تفکیک‌پذیری[۱۰] بالا را دریافت کنند و در مرحله‌ی پیاده‌سازی، محاسبات بسیار زیادی را بر روی داده‌ها انجام دهند. امروزه علاوه بر وجود رایانه‌های قدرتمندتر و تجهیزات الکترونیکی دقیق‌تر مانند اسکنرها، دوربین‌ها و صفحات رقومی‌کننده، استفاده از تکنیک‌های پردازشی مدرن و توانمند همچون شبکه‌های عصبی[۱۱] ، مدل‌های مارکوف پنهان[۱۲] ، منطق‌های مجموعهٔ فازی[۱۳] و مدل‌های پردازش زبان طبیعی[۱۴] امکان‌پذیر گشته‌است.

[۹] Semantics
[۱۰] Resolution
[۱۱] Neural Networks
[۱۲] Hidden Markoff Model
[۱۳] Fuzzy Set reasoning
[۱۴] Natural Language Processing



شکل ۱.۲: جایگاه کنونی تحقیقات در زمینهٔ سیستم های نویسه‌خوان نوری لاتین

سیستم‌های جدید نویسه‌خوان نوری برون‌خط متون چاپی و برخط متون دست‌نوشت با واژگان محدود و وابسته به نویسنده، در کاربردهای محدود به نحو کاملا رضایت‌بخشی عمل می‌کنند. اما به منظور دستیابی به هدف نهایی در شبیه‌سازی ماشینی نگارش انسانی و متون چاپی، هنوز راه درازی در پیش است. جدول ۱.۲ جایگاه کنونی پیشرفت‌های حاصل‌شده در زمینه‌ی سیستم‌های نویسه‌خوان نوری برای متون لاتین را به نمایش می‌گذارد. توجه شود که برای متون چاپی، پردازش برخط تعریف نمی‌شود.



شکل ۲.۲: برخی از ویژگی‌های نگارش زبان فارسی: الف) کلمه‌ی خورشید از سه زیرکلمه تشکیل شده؛ ب)چهار شکل مختلف حرف «ع» با توجه به موقعیت آن در کلمه، ج) هم‌پوشانی دو حرف «ح» و »ک« در کلمه‌ی «حکم»؛ د) اتصال حروف «ک» و «ا» در دو محل؛ ه) حروف متفاوت با بدنه‌ی مشابه؛ و) کشیدگی حرف «ب» در کلمه‌ی «با»

۳.۲ ویژگی‌های متون چاپی فارسی

نگارش فارسی، ویژگی‌های منحصر به فردی دارد که آن را کاملا از نگارش لاتین متمایز می‌سازد. به منظور فعالیت در حوزه‌ی نویسه‌خوان نوری فارسی، آگاهی از قوانین نگارشی و نحوه‌ی چاپ حروف در این زبان، امری ضروری است. در اینجا به ویژگی‌های کلی نگارش فارسی اشاره می‌شود.

۱. متون فارسی برخلاف متون لاتین از راست به چپ نوشته می‌شوند.

۲. در کلمات فارسی برخی از حروف از یک یا دو طرف به حروف مجاور خود اتصال دارند و برخی نیز به صورت مجزا نوشته می‌شوند. در نتیجه هر کلمه ممکن است شامل یک یا چند بخش متصل باشد که «زیرکلمه» نامیده می‌شوند. (شکل ۲.۲) چسبیده یا سرهم بودن حروف در نگارش فارسی، بازشناسی متون فارسی را برای سیستم‌های نویسه‌خوان نوری، نسبت به متون لاتین بسیار مشکل‌تر می‌سازد.

۳. حروف فارسی ممکن است چهار موقعیت مجزا و در نتیجه چهار شکل متفاوت نگارش داشته باشند: حروف ابتدایی، میانی، انتهایی و مجزا (شکل۲.۲ ب).

۴. حروف واقع در یک کلمه ممکن است همپوشانی داشته باشند، بدین معنا که نتوان با رسم خطوط عمودی، حروف را به طور کامل از یکدیگر مجزا نمود (شکل۲.۲ ج).

۵. در برخی از فونت‌ها بعضی از حروف، از یک سمت در دو محل به یکدیگر اتصال دارند (شکل۲.۲ د).

۶. برخی از حروف بین یک تا سه نقطه دارند که ممکن است در بالا یا پایین بدنه‌ی حرف واقع

باشند (شکل ۲.۲ هـ).

۷. بعضی از حروف بدنه‌ی مشابه دارند و تفاوت آن‌ها تنها در تعداد و محل قرارگیری نقاط (شکل ۲.۲هـ) یا در وجود یک سرکش است (مانند «ک»و «گ»)که در مقایسه با بدنه‌ی حروف، اندازه‌ی بسیار کوچکی دارند. این موضوع نیز یکی از مواردی است که بر پیچیدگی سیستم‌های نویسه‌خوان نوری فارسی می‌افزاید.

۸. حروف فارسی ممکن است در بالا یا پایین بدنه دارای اعراب باشند. سه اعراب ـَ ـِ ـُ در زبان فارسی، اعراب‌های اصلی‌اند و اعراب ـً در برخی کلمات عربی رایج در زبان فارسی دیده می‌شود (نظیر کلمات «عمداً و «احتمالاً»). کلمات عربی دارای اعراب ـّ و ـٍ در زبان فارسی عمومیت نیافته‌اند. هر چند کاربرد اعراب در زبان فارسی نسبت به زبان عربی بسیار محدودتر است، اما اگر کلمه‌ای نامتداول باشد یا به دلیل تشابه نگارشی آن با کلمه‌ی دیگر، تاکید بر تلفظ صحیح آن باشد، از نشانه‌های اعراب استفاده می‌شود.

۹. در بالای بدنه‌ی یک حرف ممکن است علامت تشدید وجود داشته باشد.

۱۰. برخی از حروف دارای علامت همزه هستند («ئ» ، «أ»، «ؤ»، « ۀ»).

۱۱. حروفی که از طرف چپ قابلیت اتصال به حرف مجاور خود را دارند، ممکن است به صورت کشیده نوشته شوند (شکل۲.۲ و).

۱۲. بیشتر حروف فارسی (مخصوصا حروف چسبیده) دندانه‌دار هستند. در مواردی که کیفیت سند اصلی یا دستگاه اسکنر پایین باشد، ارتفاع دندانه‌ها نسبت به خط زمینه کوتاه می‌شود و

این امر، شناسایی صحیح این حروف در مرحله‌ی قطعه‌بندی یا بازشناسی را با مشکل مواجه می‌سازد.

۴.۲ زیرکلمه

برخی حروف در زبان فارسی/عربی تنها قادرند از سمت راست به سایر حروف متصل گردند، این امر میتواند منجربه ایجاد چند بخش همبند (متصل) مجزا در کلمات گردد. این بخش‌های همبند مجزا در کلمات، زیرکلمه [۱۵] نامیده می‌شوند و به‌طور معمول توسط یک خط فرضی به نام خط پایه، به یکدیگر متصل می‌گردند. در این پروژه انواع و فراوانی زیرکلمه‌های زبان فارسی در یک متن با بیش از ۳۰۰ هزار خط مورد بررسی قرار گرفته‌است. کلمات فارسی/عربی ممکن است دو یا بیشتر از دو زیرکلمه را شامل شوند. زیرکلمات نیز معمولا از یک حرف یا تعدادی حروف تشکیل می‌شوند.

گاهی نحوه نگارشی افراد باعث ایجاد هم‌پوشی مابین زیرکلمات و هم‌بستگری دو یا چند کاراکتر می‌شود، و این امر شیوه‌های بازشناسی بر اساس تفکیک را پیچیده می‌سازد، زیرا معمولاً هیچ تفکیک‌ساز عمودی که بتواند دقیقاً حروف را از یکدیگر تفکیک سازد، وجود ندارد. لذا دو راهکار اصلی برای بازشناسی متون دست‌نویس فارسی موجود است. یک راهکار، بازشناسی مستقل از تفکیک اجزا است، که بر روی کل کلمات یا زیرکلمات اجرا می‌شود. در این روش، ابتدا کلمات را به صورت ترکیبی از کاراکترهای متصل و تفکیک‌ناپذیر در نظر می‌گیرند، سپس بااستخراج ویژگی‌هایی از کل کلمه، عمل بازشناسی اجرا می‌گردد. راهکار دیگر، بازشناسی مبتنی بر تفکیک اجزا است، که بر روی کاراکترهای اصلی سازنده کلمات یا حروف اجرا می‌شود. در این روش پس از تفکیک کلمات به کاراکترهای سازنده،

[۱۵] Stroke

شناسایی کاراکترها در سطح کلمه با استفاده از یک فرهنگ لغت اجرا شده و نهایتا با ترکیب نتایج حاصل از بازشناسی حروف، کلمات نیز بازشناسی می‌شوند.

۵.۲ مزیت استفاده از سیستم‌های نویسه‌خوان نوری

الف. افزایش چشمگیر سرعت دسترسی به اطلاعات؛ زیرا در متن بر خلاف تصویر، امکان جستجو و ویرایش وجود دارد.

ب. کاهش فضای ذخیره‌سازی؛ زیرا حجم فایل متنی استخراج‌شده از یک تصویر، معمولا بسیار کمتر از حجم خود فایل تصویری است.

چنین قابلیتی امکان استفاده‌ی گسترده از رایانه را در پردازش سریع حجم وسیعی از داده‌های مکتوب شرکت‌ها و موسسات مختلف نظیر بانک‌ها، شرکت‌های بیمه، مؤسسات خدمات عمومی، اداره‌ی پست، و دیگر نهادهایی که سالانه با میلیون‌ها مورد پرداخت، دریافت و حسابرسی امور مشتریان خود مواجه‌اند فراهم می‌آورد.

۶.۲ انواع سیستم‌های نویسه‌خوان نوری

بازشناسی متون دست‌نویس معمولا در دو حالت برخط[۱۶] (به طور هم‌زمان با نگارش متن) و برون‌خط[۱۷] (پس از نگارش متن) اجرا می‌شود. [۳] بازشناسی در سیستم‌های برخط مبتنی بر حرکت قلم (پویایی و تحرک نوشته) بوده و بر رشته‌ای از مختصات نقاط مسیر حرکت قلم در حین نگارش

[۱۶] Online

[۱۷] Offline

اعمال می‌گردد.

۱.۶.۲ سیستم‌های برخط

در بازشناسی برخط، حروف در همان زمان نگارش توسط سیستم تشخیص داده می‌شوند و دستگاه ورودی این سیستم‌ها یک قلم نوری است. در این روش علاوه بر اطلاعات مربوط به موقعیت قلم، اطلاعات زمانی مربوط به مسیر قلم نیز در اختیار است. این اطلاعات معمولا توسط یک صفحهٔ رقومی‌کننده اخذ می‌شوند. در این روش می‌توان از اطلاعات زمانی سرعت، شتاب، فشار و زمان برداشتن و گذاشتن قلم روی صفحه در بازشناسی استفاده کرد.

۲.۶.۲ سیستم‌های برون‌خط

بازشناسی در سیستم‌های برون‌خط بر تصاویر اسکن شده اسناد اعمال می‌شود و به عبارت دیگر اطلاعات پس از نوشتن، به کمک مراحل پیش‌پردازشی تفکیک کننده متن از پس‌زمینه، از تصاویر اخذ می‌گردد. [۴] پروسه‌ی نازک‌سازی و استخراج اسکلت یکی از ضروری‌ترین عملیات پیش‌پردازشی است که در بازشناسی برون‌خط به منظور حذف اثر پهنای قلم اجرا می‌شود، اما در بازشناسی برخط به دلیل اینکه قلم نگارنده خود دارای پهنای نگارشی یک پیکسل است، این عملیات اجرا نمی‌گردد. در این روش به هیچ نوع وسیله نگارش خاصی نیاز نیست و تفسیر داده‌ها مستقل از فرآیند تولید آن‌ها و تنها براساس تصویر متن صورت می گیرد. این روش به نحوه‌ی بازشناسی توسط انسان شباهت بیشتری دارد.

فصل ۳

مروری بر دادگان موجود

در این بخش، چهار پایگاه داده دست‌نویس فارسی و عربی برجسته را معرفی کرده و به بررسی ویژگی‌ها، نحوه‌ی طراحی فرم‌ها و اطلاعات آماری که در هر یک از پایگاه داده‌ها در نظر گرفته شده است، می‌پردازیم. تلاش ما براین است که از تحقیقات گذشته استفاده کرده و پایگاه داده جمع‌آوری شده ویژگی‌های بهتری نسبت به پایگاه داده‌های موجود داشته باشد.

۱.۳ پایگاه داده KHATT

KHATT[۱] ترجمه‌ی کلمه‌ی عربی «خط» به معنی دست‌نوشته می‌باشد. «خط» یک پایگاه داده دست‌نوشت برون‌خط عربی است. [۵] خط شامل ۱۰۰۰ فرم دست‌نوشت می‌باشد که توسط ۱۰۰۰ نویسنده متفاوت از کشورهای مختلف نوشته شده است. فرم‌ها با دقت ۲۰۰، ۳۰۰ و ۶۰۰ dpi اسکن

[۱] KFUPM Handwritten Arabic TexT

شده‌اند. پایگاه داده شامل ۲۰۰۰ پاراگراف تصادفی انتخاب شده از ۴۶ منبع متفاوت و ۲۰۰۰ پاراگراف مینیمال شده است که شامل تمامی اشکال حروف عربی می‌باشد. ۲۰۰۰ پاراگراف انتخاب شده‌ی تصادفی شامل ۹۳۲۷ خط می‌باشد. پایگاه داده به ٪۷۰، ٪۱۵ ، ٪۱۵ بخش به‌ترتیب برای مجموعه آموزش[۲]، تست[۳] و اعتبارسنجی[۴] تقسیم شده است. این تقسیم‌بندی محققان را قادر ساخته‌است که از این پایگاه داده برای مقایسه‌ی خروجی الگوریتم‌های خود استفاده نمایند.

۱.۱.۳ طراحی فرم

هر فرم شامل ۴ صفحه می‌باشد. در صفحه‌ی اول از نویسنده‌ها خواسته می‌شود که اطلاعتی شامل اسم، سن، کشور نویسنده، شغل، جنسیت و چپ یا راست دست بودن را پر نمایند. صفحه‌ی دوم شامل دو پاراگراف می‌باشد (یک پاراگراف ثابت و یک پاراگراف مینیمال‌شده). پاراگراف انتخاب شده واحد و در تمام فرم‌ها متفاوت می‌باشد. صفحه‌ی سوم شامل یک پاراگراف انتخاب شده‌ی متفاوت و یک پاراگراف ثابت می‌باشد که در صفحه‌ی دوم همه‌ی فرم‌ها مشترک می‌باشد. صفحه‌ی چهارم فرم‌ها برای نوشتن آزاد طراحی شده است که موجب جمع‌آوری پاراگراف‌های بیشتر با تنوع بیشتر می‌شود.

۲.۱.۳ اطلاعات آماری

داده‌های موجود در فرم‌ها از ۴۶ منبع متفاوت و ۱۱ موضوع مختلف جمع‌آوری شده‌اند. فرم‌ها توسط افرادی از ۱۸ کشور متفاوت پر شده‌اند. ۱۰۰۰ فرم توسط افرادی با سن، شغل، جنسیت و کشور متفاوت نوشته شده است.از ویژگی‌های آماری این فرم می‌توان به موارد زیر اشاره کرد:

[۲] Train

[۳] Test

[۴] Verification

- ٪۶۵ فرم‌ها توسط افراد سن بین ۱۶ تا ۲۵ نوشته شده است. دلیل این موضوع این است که بیشتر نویسنده‌ها دانش‌آموز یا دانشجو بوده‌اند. ۱۷ فرم توصط افراد بالای ۵۰ سال، ۱۲۶ فرم توسط افراد زیر ۱۵ سال و ۲۱۳ فرم توسط افراد بین ۲۶ تا ۵۰ سال پر شده‌اند.
- بیشتر از ٪۹۰ فرم‌ها توسط افراد با مدرک تحصیلی حداقل دبیرستان پر شده و تنها ٪۸.۹ نویسنده‌ها مدرک کمتر از دبیرستان داشتند.
- ۶۷۰ نویسنده از این ۱۰۰۰ نفر مرد و ۳۳۰ نفر زن بودند.
- ۸۹۱ فرم خوانا و ۱۰۹ فرم چالش برانگیز بودند.

۲.۳ پایگاه داده FHT

نحوه‌ی جمع آوری پایگاه داده FHT[۵] به این صورت بوده که نویسنده‌ها بدون استرس و در زمان کافی فرم‌ها را نوشتند. متن‌ها با سایز بزرگ چاپ شدند تا نحوه‌ی نوشتن نویسنده مانند نوشتار روزانه‌اش باشد. شرکت کننده‌ها می‌توانستند از هر ابزار نوشتاری و بدون محدودیت استفاده کنند. فرم‌های پر شده با دقت ۳۰۰ dpi اسکن شده‌اند.

۱.۲.۳ موارد استفاده

پایگاه داده FHT در موارد زیر قابل استفاده است:

- بازشناسی کلمه و زیرکلمه

[۵] An Unconstraint Farsi Handwritten Text Database

- بخش‌بندی کلمه‌ها به زیرکلمه‌ها
- تمایز بین متون چاپ شده توسط ماشین و متون دست‌نویس
- استخراج متن نوشته شده در هر خط
- تشخیص هویت نویسنده
- دسته‌بندی سند
- بازشناسی جملات فارسی

۲.۲.۳ طراحی فرم و جمع‌آوری داده

فرم‌ها توسط ۲۵۰ نویسنده در سن و تحصیلات متفاوت پر شده‌اند. ٪۶۵ از نویسنده‌ها مرد و بقیه زن بودند. از هر شرکت کننده درخواست شده است تا ۴ متن را بنویسند و هریک از متن‌ها توسط ۲۵ نویسنده‌ی متفاوت نوشته می‌شود. [۲] پایگاه داده شامل ۱۰۰۰ فرم پرشده می‌باشد که حاوی ۱۰۶۶۰۰ کلمه‌ی دست‌نوشته‌ی فارسی، ۲۳۰۱۷۵ زیرکلمه و ۸۰۵۰ جمله می‌باشد. به طور میانگین هر فرم شامل ۶.۴۵ خط، ۵..۸ جمله، ۱۰۶.۶ کلمه، ۲۳۰.۱۷۵ زیرکلمه، ۴۰۶ کاراکتر و ۱۳۲.۱ نقطه می‌باشد. هر خط به طور میانگین شامل ۱۶.۵۳ کلمه و هر کلمه شامل ۲.۱۶ زیرکلمه و ۳.۸۱ کاراکتر می‌باشد. همچنین به طور میانگین ۱۳.۲۴ کلمه یک جمله را می‌سازند.

۳.۳ پایگاه داده IfN/Farsi

در سال ۲۰۰۸ پایگاه داده IfN/Farsi شامل نام شهرها ارائه شد. [۱] این پایگاه داده شامل ۷۲۷۱ تصویر باینری است و از ۱۰۸۰ نام شهر و استان تشکیل شده است. ۶۰۰ نفر با سن و تحصیلات مختلف در این جمع‌آوری شرکت کرده‌اند. از هر نفر درخواست شده است که دو فرم شامل ۲۴ اسم شهر به همراه کد مربوطه را پر نمایند. بعد از جمع‌آوری فرم‌های پرشده اسم شهرها به همراه کد مربوطه استخراج شده‌اند.

۱.۳.۳ طراحی فرم

فرم‌ها به نحوی طراحی شده‌اند:

۱. جمع‌آوری فرم‌ها بدون محدودیت باشد.

۲. اسم شهرهای روی فرم‌ها با کیفت مشابه اسم شهرها روی نامه‌ها باشد.

۳. به آسانی پردازش شوند.

۴. فرم‌ها شامل اطلاعات اضافه راجع به افرادی که آن فرم‌ها را پر می‌کنند باشد.

هر فرم شامل سه ستون در بالا و یک متن در پایین صفحه می‌باشد. در قسمت بالایی هر فرم ۱۲ نام از شهرهای ایران و یک رقم تصادفی در خط‌های جدا پرینت شده‌اند. نام شهرها به نحوی در فرم‌ها قرار گرفته‌است که فراوانی تعداد شهرها در کل فرم‌ها باهم برابر می‌باشند. رقم‌ها نیز توزیع یکنواخت دارند. سن، شغل و جنسیت نویسنده نیز در زیر صفحه پرسیده شده است.

سن	درصد
کمتر از ۲۰	۳۰
بین ۲۰ و ۳۰	۵۰
بین ۳۰ و ۴۰	۱۰
بالای ۵۰	۱۰

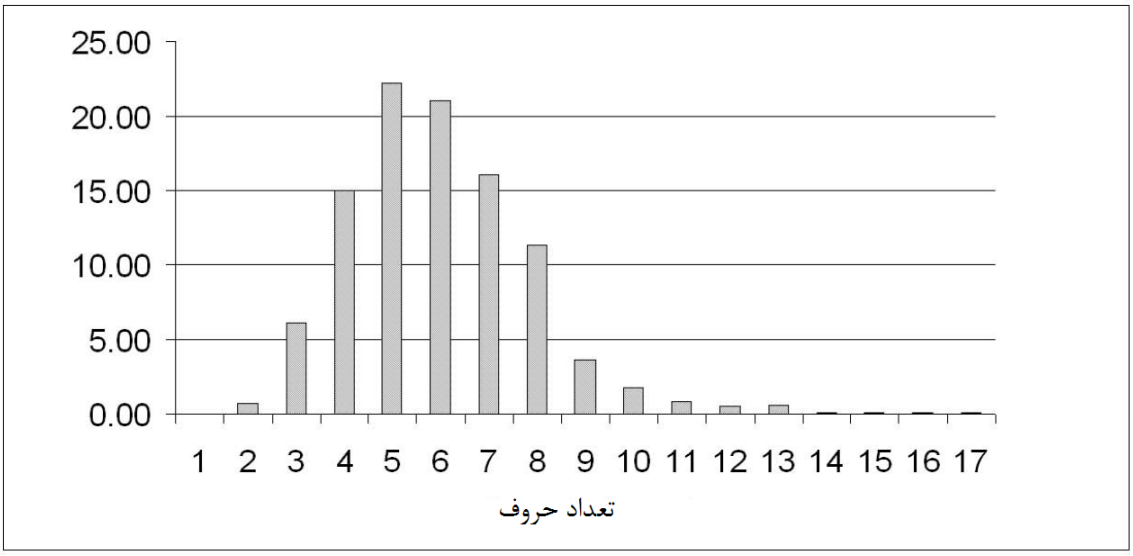
جدول ۱.۳: توزیع نویسنده‌ها از لحاظ سن

۲.۳.۳ اطلاعات آماری پایگاه داده IfN/Farsi

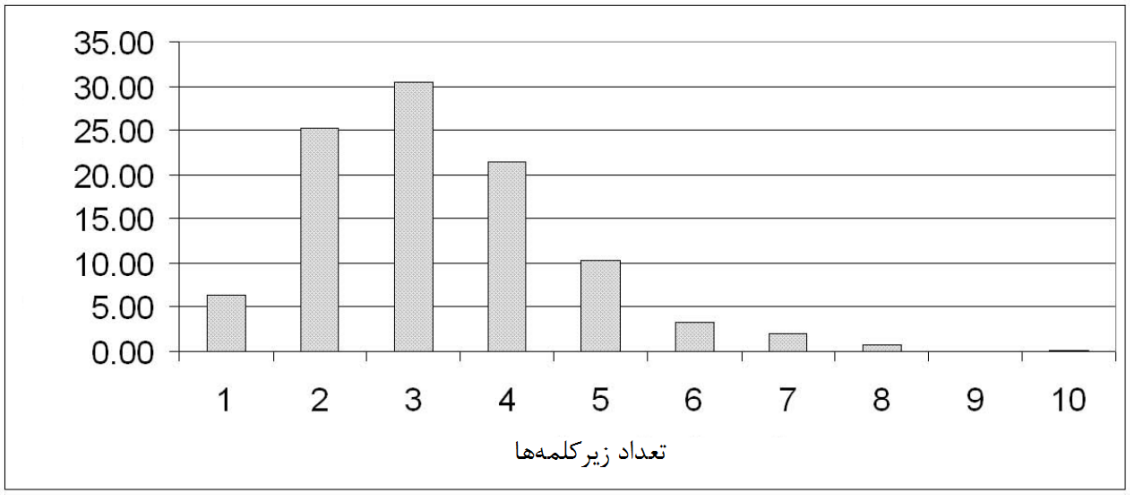
این پایگاه داده سن، تحصیلات و تجربه را عامل مهمی روی دست‌خط در نظر می‌گیرد. به همین دلیل سعی شده که نویسنده ها از سن و تحصیلات متنوعی باشند. تقریبا ٪۶۰ نویسنده‌ها مرد و ٪۴۰ آن‌ها زن بوده‌اند. جدول ۱.۳ توزیع نویسنده‌ها از لحاظ سنی را نیز نشان می‌دهد. این پایگاه داده شامل ۷۲۷۱ نمونه از ۱۰۸۰ نام شهرهای کشور می‌باشد. شکل ۱.۳ توزیع ۱۰۸۰ کلمه را بر اساس تعداد حرف‌های آن‌ها نشان می‌دهد. طبق این نمودار طولانی‌ترین کلمه شامل ۱۷ حرف می‌باشد. همچنین جدول ۲.۳ و ۳.۳ توزیع کلمه‌ها براساس تعداد زیرکلمه‌ها و نقطه‌ها را نشان می‌دهند.

۴.۳ پایگاه داده فارسا

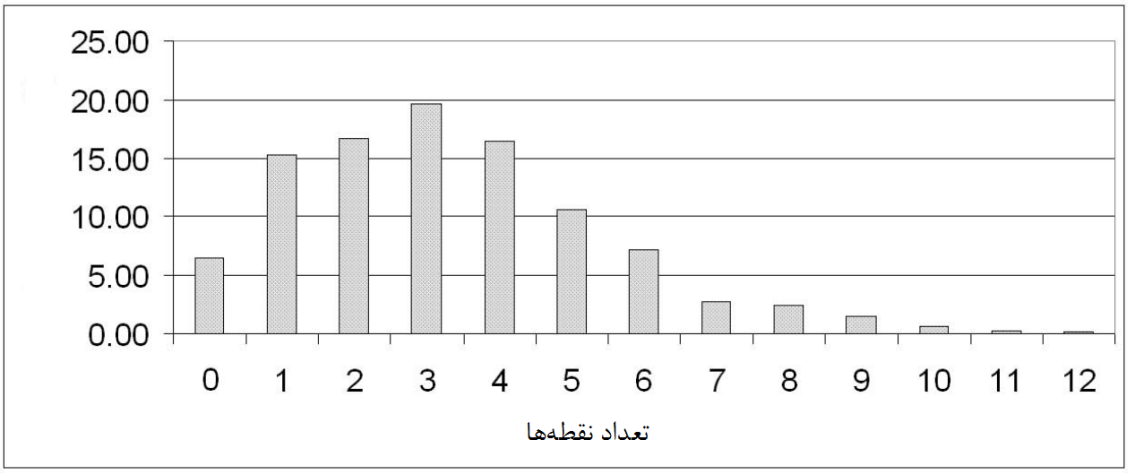
پایگاه داده فارسا شامل ۳۰۰۰ تصویر از ۳۰۰ کلمه دست‌نوشت است و توسط بیش از ۲۵۰ نویسنده از ۶۵۰ فرم مخصوص استخراج شده است و فرم‌ها با دقت ۳۰۰ dpi اسکن شده است. این ۳۰۰ کلمه یک سری از کلمات متداول پرتکرار بکار رفته در یکی از روزنامه‌های پر تیراژ کشور است. فارسا مجموعه‌ای منحصر به فرد از لحاظ تنوع دست‌نوشته از هر کلمه است. استفاده از کلمات پرکاربرد در زبان فارسی و نحوه مرتب‌سازی کلمات پرکاربرد یکی دیگر از مزیت‌های این مجموعه داده است.



شکل ۱.۳: تورزیع کلمه‌ها از لحاظ تعداد حروف



شکل ۲.۳: تورزیع کلمه‌ها از لحاظ تعداد زیرکلمه‌ها



شکل ۳.۳: تورزیع کلمه‌ها از لحاظ تعدادنقطه‌ها

هدف اصلی از این پایگاه داده فراهم آوردن شرایط مناسب برای آموزش سیستم‌های بازشناسی کلمات دست‌نویس فارسی و در نهایت کمک به خودکارسازی فرآیندهای اتوماسیون اداری و تمامی ارگان‌های مرتبط با دست‌نوشته‌های فارسی است. پایگاه داده تهیه شده نواقص پایگاه‌های داده‌ی موجود در خصوص تعداد کم کلمات را تا حد زیادی بهبود می‌بخشد. از طرفی برخلاف پایگاه داده‌های موجود که در خصوص اسم شهرهای ایران است، این پایگاه داده از کلمات پرکاربرد مطبوعات رسمی کشور تهیه شده است.

۱.۴.۳ جمع‌آوری و مرتب‌سازی کلمات پرکاربرد

روزنامه جام جم یکی از روزنامه‌های پرتیراژ ایران است. در ایجاد پایگاه فارسا یک نسخه از این روزنامه جهت استخراج کلمات مورد ارزیابی قرار گرفت. تعداد تکرار هر کلمه در متن این روزنامه

شمرده شد و کلماتی که حداقل چهار بار تکرار شده بود انتخاب شد. علاوه بر این چهار گروه دیگر از کلمات شامل:

۱. ۱۵ کلمه از اعداد

۲. سه فعل پرکاربرد: است، بود، شد

۳. حروف اضافه شامل: را، از، در، برای و یا

۴. ضمایر فاعلی: من، تو، او، ما، شما، ایشان

انتخاب شد و در کل یک مجموعه ۳۰۰تایی از کلمات لیست برداری شد.

۲.۴.۳ طراحی فرم و جمع‌آوری دست‌نوشته از افراد

به منظور تهیه‌ی پایگاه داده دست‌نوشت، کلمات استخراج‌شده در ۱۳ فرم چاپ شد. هر فرم شامل ۶ ستون و ۱۲ سطر می‌باشد. در هر فرم ۲۴ کلمه تایپ شده است. از نویسنده خواسته شد تا از هر کلمه دو بار بنویسد. اندازه پنجره‌های موجود در جدول به صورتی طراحی شده است که برای نوشتن طولانی‌ترین کلمه‌ی موجود، فضای کافی وجود داشته‌باشد. به هر نویسنده ۳عدد فرم داده شده است. درنتیجه از هر نویسنده ۱۴۴ کلمه دست‌نویس در پایگاه داده موجود است. در مجموع ۱۳ فرم مختلف ۵۰ بار تکرار شده و توسط بیش از ۱۵۰ نفر نویسنده پر شده است. به منظور همسان‌سازی نوع قلم به همه افراد یک نوع خودکار مشکی برای نوشتن داده شد.

۳.۴.۳ اطلاعات آماری پایگاه داده فارسا

پایگاه داده فارسا شامل ۳۰۰۰ تصویر از ۳۰۰ کلمه دست‌نوشته است. این تصاویر از ۶۵۰ فرم مخصوص که توسط بیش از ۲۵۰ نویسنده پر شده، استخراج شده است. در مجموع این پایگاه داده شامل ۷۴۲۰۰ زیرکلمه و ۱۳۹۲۰۰ شکل مختلف از حروف فارسی است.

فصل ۴

جمع‌آوری دادگان

در این بخش، روش‌ها و معیارهای جمع‌آوری دست نوشته‌ی افراد آورده شده است. دراین راستا، فرم‌هایی تعبیه شده تا مشخصات نویسندگان شناسایی شود و همچنین، پاراگراف‌هایی از موضوعات متنوع در اختیار افراد قرار داده شده است تا با نوشتن این پاراگراف‌ها بتوانیم مجموعه دادگانی از دست‌نوشته افراد مختلف جمع آوری کنیم.

۱.۴ دادگان جامع فارسی

برای جمع‌آوری پاراگراف‌ها،ابتدا روی یک دادگان بزرگ شامل ۳۲۹۹۹۹ خط و ۱۰۴۳۵۰۱۲ کلمه و ۸۴۶۷۹۳۶۷ کاراکتر مطالعاتی انجام شده و انواع زیرکلمه‌های این متن بزرگ پیدا شده است. طبق آمار صورت گرفته، این متن شامل ۱۱۰۰۰ زیرکلمه متفاوت می‌باشد. روش جداسازی زیرکلمه‌ها در بخش ۳.۴ توضیح داده می‌شود.

موضوع	تعداد
اجتماعی	۶۰
اقتصادی	۶۰
سیاسی	۶۰
فرهنگی و هنری	۶۰
ورزشی	۶۰
ادبیات	۶۰
علمی و دانشگاهی	۶۰
خانوادگی	۶۰
فلسفه و ریاضی	۶۰
فقه اسلامی	۶۰

جدول ۱.۴: موضوعات پاراگراف‌های جمع‌آوری شده

۲.۴ دادگان جمع‌آوری شده

تلاش ما بر این بوده است که پاراگراف‌های پیدا شده از تنوع زیرکلمه‌ای مشابه متن ذکر شده برخوردار باشد تا از جامعیت زیرکلمه‌ها اطمینان حاصل شود. پاراگراف‌های موجود در فرم‌ها ۱۰ موضوع متفاوت را پوشش می‌دهند و تنوع موضوعی موجب ایجاد تنوع در کلمه‌های نوشته‌شده در پاراگراف‌ها می‌شود. جدول ۱.۴ انواع موضوع‌ها و تعداد پاراگراف‌ها از هر موضوع را نشان می‌دهد. به عنوان نمونه تعدادی از این پاراگراف‌ها در فصل ۶ آورده شده است.

۱.۲.۴ فرم‌های آماده شده

هرفرم شامل دو پاراگراف عمومی و یک پاراگراف بهینه‌شده و ثابت می‌باشد. هر پاراگراف در دو فرم قرار داده می‌شود. یعنی ۶۰۰ تا فرم آماده شده که هر فرم شامل دو پاراگراف عمومی و یک پاراگراف

جنسیت : مرد ☐ زن ☐	سن : ۱۲-۶ ☐ ۱۸-۱۲ ☐ ۲۴-۱۸ ☐ ۳۰-۲۴ ☐ ۳۶-۳۰ ☐ ۴۲-۳۶ ☐ ۴۸-۴۲ ☐ ۴۸ به بالا ☐			
سطح تحصیلات : ابتدایی ☐ راهنمایی ☐ دبیرستان ☐ دیپلم ☐ کاردانی ☐ کارشناسی ☐ فوق لیسانس ☐ دکترا ☐ تخصص ☐		دست : راست دست ☐ چپ دست ☐	طرز نوشتار تحریری ☐ عادی ☐	
شغل : دانش‌آموز ☐ دبیر ☐ مهندس ☐ کارمند ☐ آزاد ☐ پزشک ☐		توضیحات :		

شکل ۱.۴: صفحه‌ی نخست فرم آماده شده

مشترک و ثابت می‌باشد. هر شرکت کننده دو پاراگراف می‌نویسد و هر پاراگراف توسط دو نفر نوشته می‌شود. از مزیت‌های این کار این است که اگر یکی از فرم‌ها ناخوانا بود، یک نمونه‌ی دست‌نویس دیگر نیز از آن پاراگراف موجود می‌باشد و اختلالی در جامعیت پایگاه داده ایجاد نمی‌شود. همجنین به دلیل وجود نمونه‌های بیشتر از متن‌های مشابه، دقت یادگیری ماشین افزایش می‌یابد. پاراگراف‌های جمع‌آوری شده شامل ۲۱۶۲ خط و ۶۲۲۳۹ کلمه می‌باشد. در صفحه اول این فرم‌ها از افراد خواسته می‌شود تا مشخصات خود را پر نمایند. صفحه‌ی نخست فرم مورد نظر در شکل ۱.۴ نشان داده شده است.

۳.۴ روش جداسازی زیرکلمه‌ها

برای جداسازی انواع زیرکلمه‌ها و مطالعه روی توزیع آن‌ها در متن جامع فارسی یک برنامه به زبان پایتون[۱] نوشته شده است. برنامه‌ی مربوطه در فصل ۶ آورده شده است. روند کار الگوریتم استفاده شده در این برنامه این است که ابتدا اصلاحاتی روی متن جامع انجام شده است. به این صورت که حروفی که شکل نوشتار متفاوت دارند اما تنها تفاوت آن‌ها در جای نقطه‌ها می‌باشد (حروف متفاوت با بدنه‌ی مشابه) به یک شکل درآمده‌اند. (شکل۲.۲ ه) برای مثال شکل نوشتار سه حرف «ر»، «ز» و «ژ» مشابه هم می‌باشد. بنابراین کاراکتر «ز» و «ژ» با «ر» جایگزین گردیده‌است. یعنی در جدول ۱.۴ و جدول ۴.۴ منظور از تعداد تکرار کاراکتر «ر»، تعداد تکرار هر سه کاراکتر «ر»، «ز» و «ژ» می‌باشد. برخی از کاراکترهایی که به‌طور کلی مشابه و تنها در محل قرارگیری نقاط متفاوت می‌باشند در جدول ۲.۴ آورده شده‌اند.

علاوه بر کاراکترهای آورده شده در جدول ۲.۴ دو مورد استثنا نیز وجود دارد. کاراکتر «ق» در حالتی که اول زیرکلمه و وسط زیرکلمه بیاید، شکل مشابه کاراکتر «ف» را ایجاد می‌کند. به همین دلیل در مطالعات انجام شده کاراکتر «ق» در این دو حالت خاص با کاراکتر «ف» جایگزین شده است. همچنین کاراکترهای «ن» و «ی» در حالت اول و وسط زیرکلمه مشابه کاراتر «ب» نوشته می‌شوند. بنابرین این دو کاراکتر نیز در متن اصلی در حالت اول و وسط زیرکلمه با کاراکتر «ب» جایگزین شده‌اند.

[۱] Python

حروف	حرف جایگزین
ب پ ث ت	ب
ج چ ح خ	ح
د ذ	د
ر ز ژ	ر
س ش	ش
ص ض	ص
ط ظ	ط
ع غ	ع
ک گ	ک

جدول ۲.۴: کاراکترهای با نوشتار مشابه

پس از انجام این جایگزینی‌ها الگوریتم اصلی انجام می‌شود. کلیت الگوریتم به این صورت عمل می‌کند که از ابتدای کلمه جلو می‌رود و هرجا به کاراکتر « » (فاصله)، «د»، «ذ»، «ر»، «ز»، «ژ»، «و»، «ا»، «آ»، یا خط بعد می‌رسد، یعنی یک جزء هم‌بند تمام شده و یک زیرکلمه پیدا شده است. سپس جستجو از کاراکتر بعدی کلمه ادامه می‌یابد.

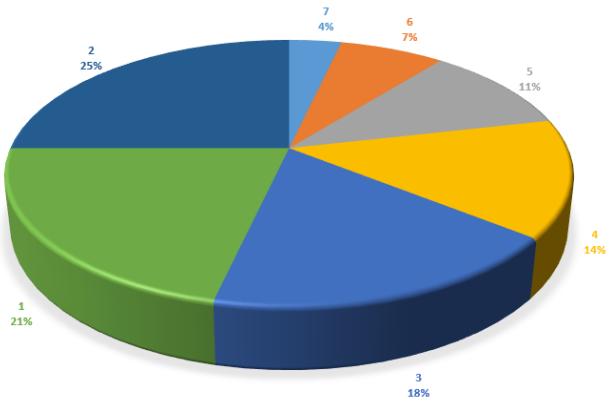
۴.۴ فراوانی زیرکلمه‌ها

بعد از اجرای روش بخش ۳.۴ روی این پاراگراف‌ها مشخص شده است که شامل ۲۴۸۱ زیرکلمه متفاوت می‌باشند.

در جدول ۴.۴ پرتکرارترین زیرکلمه‌ها در متن جامع اولیه و در جدول ۵.۴ پرتکرارترین زیرکلمه‌ها در متن جمع‌آوری شده (شامل ۶۰۰ پاراگراف) آورده شده است. همانطور که مشاهده می‌شود اکثر زیرکلمه‌های پرتکرار در متن جامع، در متن جمع‌آوری شده نیز به همان نسبت پراکنده شده‌اند.

تعداد کاراکترها	فراوانی
۱	۶۹۹۷۲
۲	۸۱۵۹۲
۳	۴۳۸۳۴
۴	۲۲۳۶۲
۵	۷۵۸۲
۶	۱۹۱۴
۷	۲۳۸

جدول ۳.۴: توزیع زیرکلمه‌ها در متن جمع‌آوری شده از لحاظ تعداد حروف



شکل ۲.۴: نمودار دایره‌ای توزیع زیرکلمه‌ها از لحاظ تعداد حروف

همچنین توزیع زیرکلمه‌ها از نظر تعداد کاراکترها در جدول ۳.۴ نشان داده شده است.

نمودار دایره‌ای ۲.۴ نیز توزیع زیرکلمه‌ها براساس طول آن‌ها را نشان می‌دهد. طبق آمارهای بدست آمده بلندترین زیرکلمه حاوی ۷ کاراکتر و کوتاه‌ترین زیرکلمه حاوی ۱ کاراکتر می‌باشد. همچنین پرتکرارترین زیرکلمه‌ها ۲ حرفی و کم‌تکرار ترین زیرکلمه‌ها ۷ حرفی می‌باشند.

زیرکلمه	تعداد	زیرکلمه	تعداد	زیرکلمه	تعداد	زیرکلمه	تعداد	زیرکلمه	تعداد
ا	۲۵۸۳۷۷۹	ما	۱۷۷۴۳۱	ببد	۶۴۹۰۹	سی	۳۵۴۱۸	کفب	۲۴۴۵۲
ر	۲۲۳۰۳۷۵	سب	۱۷۵۸۵۱	ببر	۶۰۱۶۶	ع	۳۴۴۶۳	مبر	۲۳۹۳۲
د	۱۶۵۲۰۰۶	حو	۱۷۲۳۹۲	بک	۵۸۲۶۸	بکر	۳۳۷۹۷	کد	۲۲۸۰۰
و	۱۱۶۳۴۳۲	حا	۱۴۵۵۹۳	کس	۵۶۳۸۷	مس	۳۳۱۳۲	صو	۲۲۷۹۲
ی	۶۵۵۰۹۵	ها	۱۴۳۴۴۱	لا	۵۵۵۲۷	سو	۳۲۴۵۹	بل	۲۱۹۳۹
س	۶۰۸۵۲۴	بس	۱۳۶۶۸۷	کا	۵۴۷۰۱	طر	۳۱۷۲۷	بمی	۲۱۶۷۵
با	۵۹۸۰۷۱	سا	۱۳۳۲۴۱	حد	۵۲۵۶۹	لی	۳۱۶۷۷	فد	۲۱۳۴۹
ن	۵۶۶۵۲۹	فر	۱۳۱۰۴۱	مد	۴۸۴۸۸	هم	۳۱۵۸۶	ببن	۲۱۱۳۵
بر	۴۴۶۸۵۰	ل	۱۲۴۶۳۶	سر	۴۷۹۵۸	سلا	۳۱۵۸۴	کبر	۲۰۶۷۲
به	۳۹۶۶۳۲	کر	۱۰۷۷۴۵	هد	۴۴۲۱۱	بم	۳۱۲۲۸	فو	۲۰۵۳۲
ه	۳۷۶۰۷۱	کر	۱۰۴۵۶۲	بب	۴۳۰۵۳	کبد	۳۰۷۵۵	حبا	۲۰۴۴۶
بد	۳۱۴۵۵۵	مر	۱۰۲۳۶۶	ح	۴۲۵۸۵	عر	۳۰۰۵۱	ببه	۲۰۳۴۶
آ	۲۷۸۳۹۳	مو	۱۰۰۹۴۴	ببس	۴۲۰۷۰	حبر	۳۰۰۰۴	هبد	۲۰۲۹۸
ب	۲۶۷۱۱۱	کا	۸۱۶۱۸	عا	۴۱۴۱۹	کو	۲۹۱۵۴	صا	۱۹۸۸۲
بو	۲۴۵۱۸۸	ببا	۷۷۰۶۴	بما	۳۸۶۵۹	کببد	۲۹۰۰۹	صد	۱۹۸۲۹
می	۲۱۹۷۹۳	هر	۷۱۹۴۵	مبا	۳۸۵۶۲	کو	۲۷۴۴۵	ص	۱۹۵۲۶
بن	۲۱۳۸۸۴	سبا	۶۸۲۱۵	بحا	۳۷۱۱۸	ق	۲۷۴۱۵	مبد	۱۹۴۷۷
که	۱۹۶۲۴۴	بها	۶۶۵۵۰	مه	۳۶۸۴۳	ک	۲۶۰۱۸	حب	۱۹۴۰۰
بی	۱۹۱۸۱۷	حر	۶۵۸۵۰	ف	۳۵۶۵۷	حها	۲۵۹۶۱	بهر	۱۹۲۸۱
م	۱۷۹۷۹۵	فا	۶۴۹۹۵	حه	۳۵۴۶۱	فب	۲۵۹۴۵	ببو	۱۹۰۴۹

جدول ۴.۴: فراوانی زیرکلمه‌های متن جامع

زیرکلمه	تعداد	زیرکلمه	تعداد	زیرکلمه	تعداد	زیرکلمه	تعداد	زیرکلمه	تعداد
ا	۱۶۶۹۱	حو	۱۱۷۷	سبا	۳۹۱	کی	۲۲۱	کببد	۱۴۸
ر	۱۴۳۵۱	کر	۱۰۵۰	ببر	۳۷۰	بها	۲۲۰	ببو	۱۳۷
د	۱۰۸۶۰	م	۱۰۲۸	بکر	۳۳۹	ک	۲۱۶	بل	۱۳۷
و	۷۷۹۸	ها	۹۹۱	مد	۳۲۸	حبر	۲۱۲	فب	۱۳۶
با	۴۱۷۶	حا	۸۴۴	سر	۳۱۷	صو	۲۰۵	حی	۱۳۴
ی	۴۰۷۱	می	۸۱۴	کو	۳۱۵	سی	۲۰۴	بطر	۱۳۳
س	۳۸۸۰	بس	۸۱۱	ع	۲۶۶	حه	۱۹۶	مه	۱۳۰
ن	۳۳۹۴	فر	۷۷۶	هم	۲۶۲	لی	۱۹۳	ط	۱۲۹
بر	۲۸۲۳	کا	۷۶۹	عا	۲۶۱	کبد	۱۸۴	عی	۱۲۸
به	۲۶۰۶	ل	۷۲۴	مبا	۲۶۱	بحا	۱۸۱	سلا	۱۲۸
بد	۲۴۳۶	مر	۶۷۷	حد	۲۶۰	بسبا	۱۸۰	هسببد	۱۲۸
ه	۲۱۸۳	سا	۶۵۸	هد	۲۵۷	بکا	۱۷۴	حس	۱۲۸
ب	۱۷۲۵	مو	۵۸۰	کس	۲۵۴	سو	۱۷۲	طو	۱۲۶
آ	۱۷۱۰	ببا	۵۷۶	عر	۲۵۱	ق	۱۶۴	حب	۱۲۵
بو	۱۶۴۴	ببد	۴۶۶	ح	۲۵۰	طر	۱۶۲	بفا	۱۲۴
که	۱۳۰۶	بک	۴۴۳	ف	۲۴۵	بم	۱۶۰	سبفا	۱۲۳
سب	۱۲۷۹	حر	۴۱۳	مس	۲۴۲	هبد	۱۵۲	بح	۱۲۲
بن	۱۲۷۶	لا	۴۰۴	ببس	۲۳۷	کبر	۱۵۰	کد	۱۱۹
ما	۱۱۸۹	فا	۴۰۰	بما	۲۳۶	فد	۱۴۹	که	۱۱۷
بی	۱۱۸۷	هر	۳۹۵	بب	۲۳۵	بکی	۱۴۹	حها	۱۱۶

جدول ۵.۴: فراوانی زیرکلمه‌های متن جمع‌آوری شده

فصل ۵

نتیجه‌گیری

تحقیقات زیادی در زمینه‌ی بازشناسی دست‌نوشته صورت گرفته است. اما به دلیل ضعف در وجود یک مجموعه دادگان دست‌نویس استاندارد این تحقیق‌ها هنوز به نتیجه‌ی مطلوب نرسیده و الگوریتم‌های ارائه شده دقت کمی دارند. در این گزارش یک مجموعه دادگان برون‌خط دست‌نوشت زبان فارسی جمع‌آوری ارائه می‌شود.

در فصل اول به مفاهیم اصلی و مقدماتی در زمینه‌ی پردازش الگو پرداخته‌ایم. در فصل دوم درباره‌ی ویژگی‌های متون فارسی، زیرکلمه و سیستم نویسه‌خوان نوری صحبت کرده‌ایم. فصل سوم به بررسی پایگاه‌های داده‌ی موجود و تحقیق‌های انجام شده پرداخته است. در فصل چهارم گزارش نتیجه‌ی تحقیق‌های انجام شده آورده شده، به این صورت که ۶۰۰ شرکت کننده در نظر گرفته شده که ۶۰۰ فرم را می‌نویسند. هر فرم شامل دو پاراگراف عمومی از ۶۰۰ پاراگراف جمع‌آوری شده و ۱ پاراگراف ثابت می‌باشد. هر پاراگراف عمومی در دو فرم آورده شده تا دقت یادگیری ماشین افزایش یابد.

صفحه‌ی اول فرم‌ها شامل اطلاعاتی راجع به سن، جنسیت، شغل، تحصیلات و راست یا چپ دست بودن نویسنده‌ها می‌باشد زیرا مطابق تحقیقات انجام شده این موارد روی دست‌خط افراد تاثیر می‌گذارند. از نویسنده‌ها خواسته می‌شود تا بدون استرس و مانند نوشتار روزانه‌ی خود پاراگراف‌ها را بنویسند. شرکت کننده‌ها از هر ابزار نوشتاری می‌توانند استفاده کنند.

پاراگراف‌های موجود در فرم‌ها با دقت فراوانی جمع‌آوری شده‌اند و همانطور که در فصل چهارم توضیح داده شد، فراوانی زیرکلمه‌های این پاراگراف‌ها مشابه با فراوانی زیرکلمه‌های موجود در یک مجموعه داده‌ی جامع و کامل زبان فارسی می‌باشد. بنابراین انتظار می‌رود که استفاده از این مجموعه دادگان موجب دست‌یابی به دقت خیلی بالایی شود.

فصل ۶

ضمیمه

تعدادی از پاراگراف‌های جمع‌آوری شده را به تفکیک موضوع در ادامه آورده‌ایم. در انتهای فصل نیز برنامه‌ی پایتون برای پیدا کردن تنوع زیرکلمه‌ها آورده شده است.

علمی و دانشگاهی

ناسیونالیسم در سیاست معمولا به‌عنوان یک زیرمجموعه برای دیگر باورهای همسو شناخته می‌شود و قابلیت ارتجاع به «راست و چپ» را داراست. (برای نمونه ناسیونال سوسیالیسم، یا ناسیونال دموکراسی) ناسیونالیسم شالوده‌ای برای خواست با هم زیستن واحدهای سیاسی و قومی است و متضمن این اندیشه است که فرمانروایان و شهروندان بهره‌مند از همزیستی در این واحد سیاسی فرضی متعلق به یک تبار قومی هستند. احساسات ملی ریشه در اندیشه ساخت جامعه‌ای با هویت زبانی، مذهبی، و روانشناختی مبتنی بر تصور خویشاوندی کهن اعضای یک گروه قومی فرضی است.

اجتماعی

نماینده مردم خاش در مجلس نهم با ابراز تبریک و تسلیت به جامعه فرهنگیان کشور و خانواده حمیدرضا گنگوزه‌ای ریگی، معلم شجاع و ایثارگر، خواستار اختصاص بودجه‌ای ویژه به آموزش و پرورش از محل صندوق توسعه ملی در برنامه ششم توسعه شد. با تاکید بر اینکه آموزش در سازمان حفاظت محیط زیست به معنای ارزش، سرمایه‌گذاری و توسعه نیروی انسانی است، گفت: در این سازمان محیط‌بان تنها یک نگهبان نیست بلکه یک نیروی حفاظتی تخصصی است که باید کاربرد سلاح، مقررات و شرایط اجتماعی و اکولوژیکی محیط خدمت خود را بداند.

اقتصادی

بررسی‌های میدانی نشان می‌دهد با این‌که حجم مراجعات در ماه‌های اخیر مقداری افزایش یافته، دفاتر مشاور املاک هنوز روزهای بدون مشتری را سپری می‌کنند و به گفته‌ی یکی از کارشناسان مشاور املاک در شمال پایتخت، بسیاری از دلالان از این شغل خارج شده‌اند. به گفته‌ی او روند ریزش کارشناسان مشاور املاک از سه سال گذشته آغاز شده و در سال ۹۴ شدت گرفته است. علت خروج شاغلان املاک از این بخش نیز به یک مساله مربوط می‌شود: نبود مشتری. از طرف دیگر بسیاری از سازندگان مسکن طی سه دهه گذشته سود خوبی از ارزش افزوده بخش ساختمان به جیب زدند.

سیاسی

غلامحسین دهقانی روز سه‌شنبه در سخنرانی خود در اجلاس کمیسیون خلع سلاح مجمع عمومی سازمان ملل متحد به یکی از تحولات مخرب در حوزه‌ی خلع سلاح هسته‌ای اشاره کرد و گفت: سال گذشته در نتیجه مخالفت آمریکا و انگلستان، کنفرانس بازنگری ان پی تی شکست خورد و

نتوانست به هیچ گونه توافق محتوایی در خصوص پیشبرد خلع سلاح هسته‌ای دست یابد. این دو کشور با سند نهایی کنفرانس نه بر اساس مسائل امنیت ملی خودشان بلکه صرفا به خاطر دفاع از برنامه هسته‌ای رژیم صهیونیستی مخالفت کردند.

فرهنگی و هنری

باز کردن مرموزانه و خیال پردازانه‌ی پای نویسنده‌ی داستان اصلی، البته با تغییر سرنوشت و فرجام وی، به این قصه جدید فرصتی فراهم آورده که فاصله‌ی میان دنیای دو سنت اگزوپری با بزرگ‌سالانی را که دنیای کودکانه را فراموش یا گم کرده‌اند عیان می‌کند. در مقابل می‌بینیم که با اجتناب از آفت‌هایی که برخی اقتباس‌های قلبی دچار آن بودند، اصالت پیام اصلی داستان را حفظ کرده است. از دیگر برجستگی‌های فیلم می‌توان به به‌کارگیری تکنیک‌های مختلف انیمیشن برای تفکیک میان قصه‌ی کلاسیک و قصه‌ی جدید و میان رویاها و واقعیت‌ها اشاره کرد.

ورزشی

اگر کسانی که انتقاد می‌کنند دو دوتا چهار تا داشته باشند، تا این حد نباید در مطبوعات و رسانه‌ها دست به انتقاد از تیم ملی بزنند. اگر من هم حمایت می‌کنم، به خاطر تیم ملی فوتبال کشورم و وطنم است چراکه معتقدم با نتیجه گرفتن تیم ملی اتفاقات خوب زیادی در جامعه و فوتبال ما رخ می‌دهد. اگر قرار است هر مربی در فوتبال ما کار کند، باید با اقتدار و با داشتن امکانات مناسب کار را جلو ببرد تا با کسب نتایج خوب بتواند دل میلیون‌ها ایرانی را شاد کند. . هر فردی که می‌تواند به تیم ملی کمک کند وظیفه شرعی و ملی دارد که از تیم ملی حمایت کند. تیم ملی فوتبال همانند سایر رشته‌های ورزشی نیاز به حمایت جدی دارد و با جیب خالی نمی‌توان نتیجه مطلوب را بدست آورد.

ادبیات

منظور از قالب یک شعر، شکل آرایش مصراع‌ها و نظام قافیه آرایی آن است. شعر به مفهوم عام خود نه در تعریف می‌گنجد و نه در قالب، ولی شاعران و مخاطبان آنها، به مرور زمان به تفاهم‌هایی رسیده‌اند و شکل‌هایی خاص را در مصراع‌بندی و قافیه آرایی شعر به رسمیت شناخته‌اند. در قالب‌های نوین، شاعر مقیّد نیست مصراع‌ها را وزنی یکسان ببخشد و در چیدن مصراع‌های هم قافیه، نظامی ثابت را چنان که مثلاً در غزل یا مثنوی بود رعایت کند. طول مصراع، تابع طول جمله شاعر است و قافیه نیز هرگاه شاعر لازم بداند ظاهر می‌شود. در این جا آزادی عمل بیشتر است و البته از موسیقی شعر کهن بی‌بهره است.

خانوادگی

اگرچه گاهی سرپرستی مرد تا حد تأمین هزینه‌های زندگی و موضوع تمکین خاص، تقلیل می‌یابد؛ اما واقعیت این است که سرپرستی خانواده حوزه‌های مختلف را دربر می‌گیرد. از این رو است که مرد سرپرستی و مدیریت امور فرزندان را نیز برعهده دارد و تصمیم‌گیری زن در مورد فرزندان، صرفا با اذن شوهر و به عنوان نماینده او است و به عنوان مثال، زن نمی‌تواند در اموال فرزندان بدون رضایت همسر تصرف کند. به زن توصیه شده است که کانون اقتدار خانواده را به رسمیت شناخته، شوهر خویش را تکریم کند و در غیرمعصیت خدا او را نافرمانی نکند. حتی اگر شوهر بدخلق هم باشد، زن نباید با گفتار و رفتارش او را آزرده و خشمگین نماید.

فلسفه و ریاضی

آلبرت انیشتین، یکی از بزرگترین فیزیکدانان این کره ی خاکی بود. کسی که در طول عمرش، نظریاتی مانند نسبیت عام و خاص را منتشر کرد و جهان فیزیک را دگرگون کرد. انیشتین به علت خونریزی داخلی، سال ها با بیماری‌اش دست و پنجه نرم می‌کرد. هفت سال قبل از مرگش، عمل جراحی حساسی برای پیشگیری از این بیماری انجام داد، اما متاسفانه مفید واقع نشد. این دانشمند بزرگ که در کودکی توسط معلمش خنگ و کودن لقب داده شده بود، توانست دنیا را متحول کند.

فقه اسلامی

مهم‌ترین مشکل تدوین فرهنگ فقه، کمبود متخصصان فرهنگ نویس درحوزهٔ فقه است. هرچند در حوزه‌های علمیه به ویژه حوزهٔ پر برکت قم، عالمان و فرهیختگانی که فقه را به خوبی می‌دانند و عمری را در پژوهش‌های فقهی سپری‌کرده‌اند، بسیارند، امّا آنان که با شیوه‌های فرهنگ نویسی یا دائرهٔ المعارف نویسی آشنا باشند یا به این کار رغبت نشان دهند، اندکند. گروه فرهنگ فقه برای تأمین نیروی موردنیاز خود، اقدام به جذب طلبه‌های فاضل، مشتاق و مستعد فراگیری فرهنگ نویسی‌کرد. این کار هرچند در ابتدا پیشرفت کار را با کندی روبرو ساخت!

```
import io
import re
with io.open("persian_text.txt",'r',encoding='utf8') as f:
    text = f.read()
#the main part starts after
#replacing characters with the same structure
data = re.split(' ', text)
wrd=[]
strhelp=""
s=0
c=0
help=1;
with io.open("stroke.txt",'w',encoding='utf8') as f:
    for wrd in data:
        for j in range(0,len(wrd)-1):
            if (wrd[j+1]==' ' or
                wrd[j+1]=='ر' or
                wrd[j+1]=='ز' or
                wrd[j+1]=='ژ' or
                wrd[j+1]=='د' or
                wrd[j+1]=='ذ' or
                wrd[j+1]=='و' or
                wrd[j+1]=='ا' or
                wrd[j+1]=='آ'or
                wrd[j+1]=='\n'):
                for x in range(s,j+2):
                    strhelp = strhelp+wrd[x]
                f.write('%s\n'%strhelp)
                strhelp='';
                if ((j+2)<len(wrd)):
                    s=j+2
                    j=j+2
        s=0
from collections import Counter
from string import punctuation
counter = Counter()
with open('stroke.txt') as f:
  for line in f:
    counter.update(word.strip(punctuation) for word in line.split())
result = dict(counter)
import operator
sorted_x = sorted(result.items(), key=operator.itemgetter(1))
print(sorted_x)
with io.open("dict.txt",'w',encoding='utf8') as f:
        f.write('%s\n'%sorted_x)
```

کتاب‌نامه

[۱] Mozaffari, Saeed, et al. *IfN/Farsi-Database: a database of Farsi handwritten city names.* International Conference on Frontiers in Handwriting Recognition. 2008.

[۲] Ziaratban, Majid, Karim Faez, and Fatemeh Bagheri. *FHT: An unconstraint Farsi handwritten text database.* Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on. IEEE, 2009.

[۳] Plamondon, Réjean, and Sargur N. Srihari. *Online and off-line handwriting recognition: a comprehensive survey.* IEEE Transactions on pattern analysis and machine intelligence 22.1 (2000): 63-84.

[۴] Nazif, Arica, and T. Yamin-Vural Fatos. *An overview of char-*

acter recognition based focused on off-line handwriting. IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews 31.2 (2001).

[۵] Mahmoud, Sabri A., et al. *KHATT: An open Arabic offline handwritten text database.* Pattern Recognition 47.3 (2014): 1096-1112.

[۶] Mozaffari, Saeed, and Hadi Soltanizadeh. *ICDAR 2009 handwritten Farsi/Arabic character recognition competition.* 2009 10th International Conference on Document Analysis and Recognition. IEEE, 2009.

[۷] Kherallah, M., et al. *The on/off (LMCA) dual Arabic handwriting database.* 11th International Conference on Frontiers in Handwriting Recognition (ICFHR). 2008.

College of Science
School of Mathematics, Statistics, and Computer Science

Design of Farsi Offline Handwritten Text Database

Nadia Ghobadi Pasha

Supervisor: Dr. Bagher BabaAli

A thesis submitted to Graduate Studies Office
in partial fulfillment of the requirements for the degree of
B.Sc. in
Computer Science

July 2016